Study
Note
2004-06

Comparison of Alternative Methods of Measuring ASVAB Test Composite Fairness

**Joseph Zeidner, Cecil Johnson,
Yefim Vladimirsky, and Susan Weldon**
J. Zeidner & Associates

**United States Army Research Institute
for the Behavioral and Social Sciences**

JULY 2004

**U.S. ARMY RESEARCH INSTITUTE
FOR THE BEHAVIORAL AND SOCIAL SCIENCES**

A Directorate Of The U.S. Army Human Resources Command

**ZITA M. SIMUTIS
Director**

Research accomplished under contract
for the Department of the Army

J. Zeidner & Associates

Technical Review by

Peter M. Greenston

NOTICES

DISTRIBUTION: Primary distribution of this Study Note has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-PO, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926.

FINAL DISPOSITION: This Study Note may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Study Note are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) July 2004	2. REPORT TYPE Final	3. DATES COVERED (from. . . to) Nov 2002 to Feb 2004

4. TITLE AND SUBTITLE Comparison of Alternative Methods of Measuring ASVAB Test Composite Fairness	5a. CONTRACT OR GRANT NUMBER DASW01-02-P-0355
	5b. PROGRAM ELEMENT NUMBER 665803
6. AUTHOR(S) Joseph Zeidner, Cecil Johnson, Yefim Vladimirsky, Susan Weldon	5c. PROJECT NUMBER D730
	5d. TASK NUMBER 263
	5e. WORK UNIT NUMBER C01
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) J. Zeidner & Associates 5621 Old Chester Court Bethesda, MD 20814	8. PERFORMING ORGANIZATION REPORT NUMBER
9.SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) U.S Army Research Institute for the Social and Behavioral Sciences 2511 Jefferson Davis Highway Arlington, VA 22202	10.MONITOR ACRONYM ARI
	11. MONITOR REPORT NUMBER Study Note 2004-06

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT *(Maximum 200 words)*

The major objective of the present research is to compare the fairness measures obtained by the prediction error (PE) model with the Cleary model for female and black Soldiers utilizing the current Army aptitude area composites as predictors and Skill Qualifications Test (SQT) scores as the criteria. Fairness is traditionally defined as the absence of underpredictions for the minority groups that are considered potentially susceptible to discrimination. The Cleary model was chosen for comparison with the PE model because Cleary has been considered the "gold standard" of fairness measurement for more than three decades.

The models are compared for selection and classification and evaluated by a common metric using the same robust Army database in a double cross-validation design permitting objective estimates of prediction fairness.

The authors conclude that the results obtained for the PE and Cleary models are quite comparable for practical purposes for selection, but possibly not for classification. They find that the PE model is the better of the two because of the precision that comes with utilization and reliance upon individual test scores.

15. SUBJECT TERMS
Armed Services Vocational Aptitude Battery (ASVAB), Army aptitude area composites, personnel classification, fairness of performance prediction tests for minorities, optimal classification and assignment

SECURITY CLASSIFICATION OF			19. LIMITATION OF	20. NUMBER	21. RESPONSIBLE PERSON
16. REPORT	ABSTRACT	18. THIS PAGE	ABSTRACT	OF PAGES	(Name and Telephone Number)
Unclassified	Unclassified	Unclassified	Unclassified	70	Dr. Peter Greenston 703-602-7944

Study Note 2004-06

Comparison of Alternative Methods of Measuring ASVAB Test Composite Fairness

Joseph Zeidner, Cecil D. Johnson,
Yefim Vladimirsky, and Susan Weldon
J. Zeidner & Associates

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U. S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

July 2004

Army Project Number
2O465803D730

Personnel and Training
Analysis Activities

Approved for public release; distribution is unlimited.

ACKNOWLEDGEMENT

The authors are grateful to Dr. Peter Greenston of the Army Research Institute for his contribution to the research effort, including his support and encouragement, and careful review and editing of the draft report.

Comparison of Alternative Methods of Measuring ASVAB Test Composite Fairness

Requirement

At an individual or group level, the difference between predicted performance and actual performance can be defined as prediction error in a model of performance. Fairness is traditionally defined as the absence of underpredictions for the minority groups that are considered potentially susceptible to discrimination. A major objective of the present research is to provide evidence that the Prediction Error (PE) model, in which current Army aptitude area (AA) composites are used to predict performance, measures fairness as well as the traditional Cleary measure of fairness. The specific goal is to compare the results of measuring fairness in each of these two models using an Army job performance database covering a wide variety of MOS. The research employs a common metric across both models, focusing on MOS for which PE differences were previously found to be statistically significant.

Procedures

The PE method uses the MOS total group regression parameters to estimate predicted scores, and applies these total group parameters in calculating predicted versus actual differences for minority subgroups. The Cleary fairness measure depends on differences between regression lines using parameters computed in the total sample and in the minority group samples.

In previous research, ARI's ASVAB and Skill Qualifications Test (SQT) database for FY 1983-1989 was used to obtain a sample of about 83,000 Soldiers in 66 MOS. Twenty-six of these MOS – comprising 36,000 Soldiers, of which about 12,000 were females or blacks – were subsequently selected for a more intensive study of fairness in the present research. These 26

MOS showed statistically significant performance underpredictions for female and/or black subgroups.

Findings

There were a large number of individuals underpredicted by both models using the common metric. Using the Cleary measure, 50 percent or more of females were underpredicted in 19 of 25 MOS, and 50 percent of more of blacks were underpredicted in 13 of 25 MOS. The most severe underpredictions of fairness were for females performing traditional work in the administrative and clerical areas. In testing for statistical significance using t-tests for the PE model, 16 of 25 MOS for females and 9 of 25 for blacks were found significant at the .05 level or higher. The correlation between the PE fairness model and Cleary model across the 25 MOS was .95 for females and .90 for blacks.

Conclusions

The authors found that overall the two models could be considered comparable as measures of fairness, while preferring the greater precision of the PE model with its reliance upon individual scores. Because the prediction error differences were, in general, small, they were not of practical significance in selection or for setting minimum enlistment scores (i.e., cutoff scores). The same differences, however, might well have implications for classification decisions within a classification optimization framework.

Comparison of Alternative Methods of Measuring ASVAB Test Composite Fairness

Table of Contents

INTRODUCTION

Background

This report describes three methods that may be used to determine the fairness of the ASVAB test composites for 26 out of 66 MOS previously shown to have statistically significant prediction error scores by gender and race (Zeidner, Johnson, Vladimirsky and Weldon, 2004). The methods compared are the Prediction Error method (Zeidner, Johnson, Vladimirsky and Weldon, September, 1998 and Zeidner et al., 2004), the Cleary method or Regression Model (Cleary, 1968), and the Thorndike method or Constant Ratio Model (Thorndike, 1971). The three definitions of prediction fairness differ in several important ways.

Fairness is traditionally defined as the absence of underpredictions for the minority group for which discrimination potentially exists (Cleary, 1968). Thus, if a test is used for selection and is underpredicting minority group performance, members of a minority group may be rejected for a job that they were capable of performing successfully.

For the PE method, fairness is defined as the difference between predicted Skill Qualification Test (SQT) scores using the regression weights computed in the total group and actual SQT scores for the total group. These differences are computed for female or black subgroups within the same MOS, depending on whether racial or gender bias is being investigated. Complete fairness is indicated by very small differences, and fairness to minorities is present when mean difference in the minority group is zero or has a positive sign (overprediction). For the Clearly model, differences in total and minority group regression equations are used to measure fairness. Thorndike (1971) uses a modified model of fairness that holds a selection measure is fair only if the success ratio for a specified criterion equals the selection ratio. Cascio (1991) gives an example: "if 40% of the minority group members are successful and 70% of the non-minority group members are successful, the proportion of minority group members selected should match the 40:70 success ratio" (p 183).

In this study fairness is measured at the individual level and then aggregated to the MOS level. This is readily accomplished with the PE method where the basic measure is a difference computed at the individual level. While the Cleary method is usually described in terms of the distance between two regression lines, this difference is equal to the aggregation of differences between pairs of predicted scores at the individual level. The Thorndike method calls for a comparison of two numbers computed for a group; one of these numbers can be subtracted from the other to provide a scale that is zero at the point of perfect fairness. This scale can be further adapted to provide fairness scores for individuals (i.e., differences between the above two numbers when aggregated over a group). It is this adapted scale for individuals that could be utilized in the Thorndike method and compared to the PE and Cleary methods.

<u>The need for this study arises from concern that the traditional models used to determine fairness evolved in the context of selection.</u> In contrast, Zeidner et al. (1998, 2004) employ a classification approach that uses the full range of test composite scores. Moreover, there is a need to examine the general findings in fairness studies showing overpredictions for minority groups as contrasted to the earlier findings of our PE method that consistently show more underpredictions than overpredictions for females and blacks (Zeidner et al., 1998, 2004). In the present study, the PE and Cleary models are compared for classification and evaluated by a common metric on the same robust database in a double cross-validation design permitting unbiased estimates of the PE fairness measures. A double cross validation design was not used

for the original Cleary measures since only back sample designs were described in the published studies.[1]

The focus of all fairness models is, generally, to reliably and ethically differentiate between actual and predicted job performance for all groups of individuals. Acceptability is a continuing problem of consequence for fairness measures that use different standards or separate regression equations for minority and total groups. The military, government agencies, and business organizations would most likely find such methods unacceptable. Despite the goal to improve job performance through the use of tests, social policy, employment law and ethical considerations make it essential not to discriminate against minority groups in making personnel decisions.

Cascio (1991) describes in some detail five fairness models (from about a dozen) concluding that "there is more than one reasonable definition of selection fairness and the definitions have different practical and ethical implications that may conflict. Moreover, these are irreconcilable among various ethical positions. . . ." (p. 185). Cascio, Outzz, Zedeck and Goldstein (1991) make the same point with regard to the use of test bands or intervals. Guion (1991, 1998) provides an excellent discussion of bias and fairness distinctions and problems. Guion (1998) points out that "test bias is a psychometric term referring to distortion from different unwanted sources of variance in scores from different groups. Adverse impact is a social, political, or legal term referring to an effect of test use" (p. 442). Adverse impact, Guion writes, occurs for a number of reasons and proceeds to list six of them, including bias. However,

[1] In Cleary's seminal publication of 1968, she evaluated black and white college student groups, regardless of gender. At the end of the first year of college, she obtained SATs (the predictors) and GPA (the criteria). Students were at three state supported colleges. The students were separately analyzed for each college. Blacks totaled 273 and there were over 2,000 white students. The authors of the present study believe that it is fair to say that the minority students were very carefully selected on the basis of grades, motivation, and other relevant factors. In this academic context, Cleary found that in two of the colleges there were not significant differences in the regression lines for blacks and whites. In one college, however, blacks were overpredicted using the common regression line.

he points out that it would be necessary to rule out the other five before accepting bias only as the cause. Guion suggests a variety of statistical tools for bias analysis including analysis of variance, factor analysis of various types, and differential item functioning. He also cautions that in criterion-related validation, we need to insure that the criterion is free from third-variable biases and also emphasizes how difficult it is to accomplish this goal.

Guion (1998) goes on to write that "A special issue of the *Journal of Education Measure* (Jaeger, ed., 1976) may have stilled debate over the models; it demonstrated the futility of looking to statistical models to answer political or social questions. Most participating authors looked to more rational, explicit values and the development of decision algorithms to maximize both organizational and social utilities" (p. 442). Clearly the fairness models are models of test use, not models of bias inherent in test scores.

Gottfredson (1998), writing on the fairness of tests, points to the crux of the issue:

> The vulnerability of tests is due less to their limitations for measuring important differences than it is their very ability to do so.... Keeping the spotlight on tests merely forestalls the real debate – can this society justly and constructively deal with the racial and ethnic differences in ability that will be with us for some time to come? (p. 294).

Prediction Fairness in Military Studies

McLaughlin, Rossmeissl, Wise, Brandt, et al. (1984) basically followed the Cleary (1968) method of examining fairness using regression lines. However, they compare black regression lines against white lines and female regression lines against male lines (rather than total group against minority group lines as Cleary does). Even though "the Army does not use separate black and white regression lines to select and classify enlisted personnel," they point out that "the relation becomes important when significant differences between the subgroup lines exist" (p. 60). They go on to assert that a "ten-point difference in their composites [when comparing two individuals] would not affect either person's selection or classification" (p.60). However, McLaughlin et al. are really not addressing the classification problem. It is clear that the aptitude area composites were being used only as an additional selection procedure at that time and for establishing minimum cutoff scores for entrance into training programs; in fact, the AFQT (Armed Forces Qualification Test) is used to select recruits in the youth population for military service, and most such cutoff scores were set so low as to have little effect.

Wise, Welsh, Grafton, Foley, et al. (1992), examining the technical composites of the ASVAB, found that whites had statistically significantly higher expected criterion scores than blacks in the military services. The authors state that while the differences are of statistical significance (in these large samples), they are of limited practical significance, being only about one-tenth of a standard deviation. The overall results also showed that males had higher expected criterion scores than females (except at the highest level of the selection composite scores).

The McLaughlin et al. study found, in general, significant overpredictions for blacks that were consistent with the Wise et al., 1992, study, but not in accord with the Zeidner et al., 1998, 2004 findings showing more underpredictions of MOS using least-squares estimation (LSE) composites. McLaughlin et al. did find three aptitude area composites – clerical (CL), operators / food (OF), and surveillance / communication (SC) – with significant underpredictions for females, findings that are generally consistent with the Wise et al., 1992, and Zeidner et al., 1998, 2004 studies. The CL composite fairness underprediction for females in McLaughlin et al. had the largest difference among all the aptitude areas, but the authors conclude that the use of the aptitude area composites would not result in unfair practices against blacks or females. It is interesting to note that both the McLaughlin et al. and our earlier PE studies found significant underpredictions for females in the aptitude areas corresponding to the most traditional area of work: clerical and administrative.

In evaluating prediction error scores (PEs) resulting from operational assignment to MOS and job families, Zeidner et al. (2004) found a distinct pattern of underpredictions for blacks and females. In testing for statistical significance, eleven PE means were found to have statistically significant differences from zero at the .05-level for the set of 66 MOS. For females, 17 MOS had statistically significant underpredictions among the 50 MOS containing females. Perhaps more importantly, in testing mean differences, prediction error differences for blacks and females were too small to have practical significance for selection. For blacks, the overall mean prediction error was -.019 of a standard deviation across the 66 MOS or .38 in Army aptitude area (AA) standard score units. (Aptitude areas have a mean of 100 and a SD of 20 in the youth population.) For females, the mean prediction error was -.108 of a standard deviation or 2.16 in AA standard score units.

These fairness findings for minorities are consistent with research findings in both civilian employment and military settings that are depicted in terms of regression line differences. These differences were primarily due to differences in intercept values. Such differences were also found in Zeidner et al. (1998). Lower test scores for minorities appear to be a relatively common phenomenon. In Zeidner et al., minority groups were underpredicted when prediction errors were based on the LSE test weights estimated across all individuals, as is appropriate for an operational military system.

The overall conclusion in the Zeidner et al. 1998, 2004 studies, then, was that the LSE composites (adopted in January 2002) provided substantial improvements in classification efficiency over the earlier unit-weighted composites, with little practical consequence for the selection process as a result of underpredictions found for minorities. Other distinctions among the models are noted below and in the Methods section.

General Approach

The primary objective of this study is to compare the PE and Cleary methods using a common database for assessing the effectiveness of fairness measures. The results already provided by our PE method from previous fairness study reports (Zeidner, et al., 1998; 2004) are expanded to include the results from the Cleary method. A common metric, referred to as CM, is used in conjunction with both the PE and Cleary fairness methods, along with "t-tests" computed for the differences of the PE fairness means in each minority group from the PE fairness means in the total sample of each MOS. The Thorndike method for determining racial and gender fairness will be discussed as an alternative to the PE and Cleary methods but will not be evaluated in the present study.

Comparison of Prediction Fairness Methods

The three models of fairness result in different indices of fairness because of differences in definitions of fairness and methods of computing both the predictor and criterion variables used to implement these definitions. Cleary's definition of fairness is:

> A test is biased for members of a subgroup of the population if, in the prediction of a criterion for which the test was designed, consistent nonzero errors of prediction are made for members of the subgroup. In other words, the test is biased if the criterion score predicted from the common regression line is consistently too high or too low for members of the subgroup. With this definition of bias, there may be a connotation of "unfair," particularly if the use of the test produces a prediction that is too low. If the test is used for selection, members of a subgroup may be rejected when they were capable of adequate performance (p. 115).

Schmidt and Hunter (1974) write that the Cleary definition came to be accepted by most working in this area. In Guion's (1998) view, "the Cleary model has generally been accepted as perhaps the best of a poor set of choices" (p. 442).

Thorndike's (1971) definition of fairness contends that a test is fair only if, for a specified level of criterion performance, the selection measure provides the same proportion of minority applicants that would be selected by the criterion itself. Thorndike holds that when two groups differ in mean test score, the test may be unfair to the lower scoring group as a whole "in the sense that the proportion qualified on the test may be smaller, relative to the higher scoring group, that will reach every specified level of performance" (p. 63).

We note that our PE method in the present research appears to differ from the above definition of Cleary's method. The PE method uses the MOS total group regression parameters to estimate predicted scores, finds the difference between these predicted and the actual criterion scores, and applies these total group parameters in calculating predicted versus actual differences for minority subgroups. In contrast, Cleary's method finds the difference between the MOS total prediction (or regression line) and the prediction (or regression line) computed within the minority group. In effect, she obtains the difference between the predicted performance of an individual considering the individual to be a member of the minority subgroup and a second predicted performance considering the individual to be a member of the total group. Both predicted performance scores are least square estimates of the criterion based on "composite" scores.

Discussion of Fairness Concepts

As noted, in general, when predictor scores are larger than criterion scores, the condition of overprediction exists; when predictor scores are smaller than criterion scores, the condition of underprediction exists. For a given minority individual with a vector of test scores, the prediction based on total group parameters yields a higher predicted score than does the prediction based on black / female group parameters (because parameters computed in the total sample are larger than those computed in a minority sample). The Cleary fairness measure is equal to the difference between these two predicted scores, and can be expected to be smaller when there is more overlap between the two samples on which the parameters are computed. Intuitively, this difference between the two predictors will be smaller when the prediction based on the MOS total sample has a greater proportion of blacks or females (as with the Army as contrasted to the other services). Thus, one can expect to find more underprediction based on the Cleary fairness measure using a total group predictor that has a higher proportion of black or female individuals than using a predictor based on a lower proportion of black or female individuals (e.g., the other services).

Also, many of the earlier studies on fairness focused on the lower end of the regression line – since their concern was with initial selection, rather than with classification. Obviously selection occurs at the low end of the regression line. In contrast, classification occurs over the entire distribution of predictor scores. The methodology of these earlier investigators may often find overpredictions among the low scorers where the standard error of measurement may differ from those individuals found higher on the predictor distribution. Finally, the present authors believe that the use of prediction error scores (the PE method) more precisely measures the degree of over- or underpredictions rather than just observing the occurrence of over- and underpredictions through comparison of regression equations.

OBJECTIVES

The present study is designed to obtain fairness scores for several methods across 26 MOS using composites based on seven LSE-weighted ASVAB tests. While gender and race data were available for 66 MOS and 9 job families, only 26 MOS had statistically significant fairness results for minority groups in the earlier study (Zeidner, Johnson et al., 2004). The 66 MOS parent database was selected as being the only robust data set extant containing race and gender information for individuals along with other required variables available to the researchers. The specific objectives are:

1. To compare the results of measuring gender and racial fairness using the PE and Cleary fairness measures and an associated common metric for each model, using MOS previously found to provide statistically significant results for the PE fairness measure in the context of classification (i.e., using the full range of test scores).
2. To indicate the importance of the findings to the choice of test and test composites for use in future operational systems of the selection, classification and assignment of recruits to the Army.

METHOD

Description of Data: SQT

Prior to 1983, the Skill Qualifications Test (SQT) had both written and hands-on components measuring job proficiency. After 1983, the SQT was designed only as a task-based written test of job proficiency. Soldiers were required to take the SQT annually after completing 11 months or more of service.[2]

[2] In a related study, SQTs were found to be equivalent to specially developed hands-on performance measures used as criteria in Project A (Zeidner, Scholarios and Johnson, 2003). Equivalence was defined as making the same decisions or as having similar outcomes employing either criterion.

ARI's database for SQT years FY 1987 – 1989 was used to obtain a sample of 83,132 Skill Level 1 soldiers in 66 MOS. At the time of this study, those years were considered by ARI to be psychometrically good SQT years in terms of discriminability and reliability of the measures.

As mentioned, twenty-six (of these 66) MOS were selected on the basis of earlier results (Zeidner, Johnson, et al., 2004) for the current study. Those selected had critical ratios of 1.99 or higher for the PE fairness score differences between black and white groups, or male and female groups. McNemar (1949, pp 63-66) describes critical ratios of the type used to make this selection. As shown in Table 1, the sample size for the total of these 26 MOS is 36,328.

A double cross-validation design permitting complete unbiased estimates of prediction fairness is used. Each MOS sample is divided randomly into half samples for use in a double cross-validation design. The criterion variable is corrected for attenuation, and standardized to have a mean of zero and a standard deviation of one within a single MOS. Regression weights are computed in one half of each MOS sample and applied to scores in the other half to obtain least square estimates of the criterion. The results from the two cross samples are then aggregated.

Composite Scores

The predictor scores used for both the Cleary (and Thorndike) methods are a function of all 7 ASVAB tests, computed separately for each MOS, and referred to as composite scores. "Composites" are the sum of "best" weighted ASVAB test scores that have been converted to Army standard scores in the youth population (YP).[3] The "best" weights used in creating these composites are least square regression weights for predicting SQT scores. These composites are

[3] See Appendix A for details of the conversion process.

used as the single predictor for computing black, female, and MOS total group prediction scores in the Cleary fairness measures.

The least square estimates of the SQT criterion, computed as a function of the composite scores for the Cleary measure and proposed Thorndike adaptation, are in the form $y = bx + c$, where y is a composite score computed by applying a weight to each operational test score (x) and adding a constant. These converted best weighted test scores are calculated to provide Army standard scores (mean of 100 and SD of 20) in the youth population (YP), and the b weights applied to these converted test scores are equal to the validity of the composite scores multiplied by the SD of the criterion and divided by the SD of the composite scores, after correcting the validity coefficient and both SDs for restriction in range to the YP. LSE parameters, b and c, are computed for the MOS total group, and for each race and gender subgroup.

Prediction Fairness Measures

The prediction error (PE) method fairness score for each individual equals the total group LSE score for that individual minus his or her criterion score, after both scores are converted to statistical standard scores (separately for each MOS). The PE method computes a best fitting regression line based on the 7 test scores, enters the individual's test scores into this function to obtain a predicted performance (PP) score, and then converts this PP score into a statistical standard score (a mean of zero and a standard deviation of 1.0 within the total MOS). This makes both of the variables used to compute the PE fairness measure into statistical standard scores. A negative (positive) difference score indicates underprediction (overprediction) for the individual, and a negative (positive) mean difference indicates underprediction (overprediction) for the group.

A Cleary fairness score for each individual equals the difference between a LSE score computed in the minority group of a MOS and the LSE score computed in the total MOS group. The mean of these difference scores computed for a minority group is the conventional Cleary gender or race fairness score for the MOS. This mean is an algebraic equivalent to the average distance between the regression line that predicts the criterion in the minority group and the corresponding regression line for the total group. A negative difference score indicates underprediction for an individual and a negative mean of these differences indicates underprediction for the minority group. Positive differences similarly indicate overprediction, and zero differences indicate complete fairness.[4]

Correction of Validities

The validities in this study were corrected for restriction in range separately by MOS. Range restriction was due to operational assignment effects, the restriction in range impact of assignment to MOS from a common entry pool (see Appendix B). Since the PE method of the present study uses the Army input sample rather than the youth population as the basis for making this correction, no further correction is made for restriction due to selection effects; in the Cleary (and Thorndike) methods, validities are additionally corrected for selection into the

[4] The Thorndike measure of prediction fairness is computed by first obtaining the number of individuals having predictor scores greater than a specified predictor cut score. Then the number of individuals having criterion scores greater than the LSE of the criterion score corresponding to the "specified predictor score" is subtracted from the first number. Complete fairness is indicated when this difference is equal to zero. This aggregated measure shows underprediction when negative and overprediction when positive. The individual scores for this Thorndike measure are equal to –1, 0, or +1.

A score for each individual in black and female subgroups of each MOS that, when summed over the group, becomes the Thorndike fairness is computed by using the following procedure. First, a Thorndike predictor value (TPV) score is obtained by subtracting 100 (our designated cut score) from each individual's composite score and convert all negative differences to –1, zero differences to 0, and positive differences to +1. Similarly we obtain a Thorndike criterion value (TCV) score by subtracting the total group LSE, a f(x) using x equal to the predictor cut score, from the individuals MOS criterion score, and again converting all positive scores to 1.0, all negative scores to –1, and all zero scores to 0. The Thorndike fairness score (TFS) can now be computed for each individual; TFS = TPV minus TCV. Each individual will have a –1, a 0, or a + 1 as a TFS score. Note that a score of zero which indicates perfect fairness for PE and Cleary fairness scores may indicate either perfect fairness or no information on fairness for the Thorndike fairness measure.

Army. Validities are also corrected for unreliability of the criterion variable prior to the restriction in range correction. For these and other scaling differences, the three methods cannot be compared directly with each other, hence the use of a common metric.

A Common Metric

An effective comparison of the PE and Cleary methods requires the use of a common metric, since the scales of the two measures are not equal. A useful metric to this end is one for which the magnitude of over- and underprediction can be compared across the three methods. Each common metric (CM) measuring prediction fairness is computed separately for both fairness measures, and separately for black and female groups within each of the 26 MOS. The CM is defined as the number of individuals in a minority group that are underpredicted by the PE or Cleary fairness measures. This is simply the number of negative fairness measure scores in each of the minority subgroups for each MOS. For the Thorndike fairness measure, this is the number of individual fairness scores equal to –1.

A Test of Statistical Significance

Both the PE and Cleary methods in this study, as well as would be the case for the Thorndike method, are zero at the point of perfect fairness for both the individual and group level. For each of these fairness methods a "t-test" can be computed for each job family reflecting the significance of the difference between the mean fairness measure computed in a minority sample of a MOS and a mean fairness measure assumed to be the population value for that MOS. The mean fairness measure in the total sample is our best estimate for this population value. These "t-tests" are computed only for the PE fairness measures.[5]

[5] We did not include t-test values for the CFM in Tables 3 and 4 because these fairness measures are partially a function of group membership, in contrast to the PE measures that are obtained independently of knowledge of group membership. In other words, Cleary t-tests would have been biased towards higher values. We, however,

A separate t-test score is computed for female and black sub-samples of each of the 26 MOS. These scores can be used to estimate the statistical significance of the differences of PE fairness score means in each minority sub-sample from the population. This population mean is estimated using the mean fairness measure in the MOS total sample. The t-test score is computed as a function of the minority group sample size (N), standard deviation (SD) of the measure in the minority group, and the difference of the minority group and total means. The simplified formula utilized to compute these t-test values for each MOS is as follows: $CR = [(N)^{\frac{1}{2}}]$ [(difference between means) / SD] (Ostle & Mensing, 1975). It should be noted that these t-test values differ from the CRs used to select the 26 MOS (for closer inspection in this study) from the larger set of 66 MOS.[6]

(continued) include in Appendix C the nature of the bias that consists of overlapping information in CFM scores and subgroup means. This overlap, in turn, inflates both the correlation between CFM and group membership and the size of the t-tests. The effects of the inflation of the t-test due to the use of ingredients that are inflated (by being computed in the same sample as regression line parameters are computed) are an additional contamination. Both reasons make the t-tests for the Cleary model unsuitable for use as significance tests. We, however, show the inflated Cleary t-test results in Appendix D. Note, also, the CFM for a minority group decreases as the ratio of the minority group size to the total group size decreases as shown in Appendix E.

[6] The critical ratios described in the earlier report are for the differences between fairness means for black and white groups, and between male and female groups, while the t-tests used to test the significance of the mean PE fairness are for the differences between a minority group and the total group.

RESULTS AND DISCUSSION

Description of MOS

Table 1 lists the 26 MOS by number and name used in this study. As noted earlier, these MOS were selected from a larger sample of soldiers in 66 MOS on the basis of having significant critical ratios of 1.99 or higher for the PE differences between male and female groups or white and black groups in an earlier study (Zeidner et al., 2004). Table 1 also shows the N for each sample and the total N of 36,328. The average sample size of the 26 MOS is 1337, although a couple of MOS are small, e.g., 33T has an N = 71 and 81E has an N = 129. For most MOS, the size of minority group Ns range from 30 to 2015.

Prediction Error Measured for 26 MOS for the PE Method

Table 2 shows the mean prediction error (PE) scores or PE fairness measures in each of the 26 MOS. It also shows the t-test and common metric scores (CM). CM shows the number of minority individuals in each MOS that are underpredicted as a percentage.

At the end of Table 2 are shown the grand overall arithmetic and absolute mean PE fairness measure. Also shown are the number of mean over- and underpredictions by subgroup.

Table 1

Number of First-Term Enlistees Assigned to Each of 26 MOS by Gender and Race in the FY 1987 - 1989 Data Set

		Percent				
MOS	Name	Male	Female	White	Black	N
13M	Multiple Launch Rocket Sys (MLRS) Crewmember	100.00	0.00	91.20	8.80	375
16D	Hawk Missile Crewmember	88.53	11.47	82.44	17.56	279
31K	Combat Signaler	92.04	7.96	60.62	39.38	2,750
31V	Unit Level Communications Maintainer	92.48	7.52	70.79	29.21	1,729
33T	EW/I Tactical Systems Repairer	95.77	4.23	98.59	1.41	71
45K	Tank Turret Repairer	97.87	2.13	83.54	16.46	328
52D	Power Generator Equipment Repairer	95.45	4.55	78.20	21.80	2,394
54B	Chemical Operations Specialist	92.30	7.70	72.45	27.55	1,078
55B	Ammunitions Specialist	91.40	8.60	72.25	27.75	919
63B	Light-Wheel Vehicle Mechanic	91.01	8.99	73.98	26.02	4,439
68J	Aircraft Armament/Missile Systems Repairer	96.73	3.27	83.65	16.35	367
71D	Legal Specialist	68.73	31.27	80.00	20.00	550
71L	Administrative Specialist	31.11	68.89	46.80	53.20	765
71M	Chaplain Assistant	66.05	33.95	79.05	20.95	377
72E	Tactical Telecommunications Ctr Op	78.68	21.32	58.31	41.69	638
72G	Automatic Data Telecommunications Ctr Op	49.92	50.08	59.78	40.22	649
73C	Finance Specialist	56.20	43.80	54.69	45.31	799
75B	Personnel Administration Specialist	68.16	31.84	53.24	46.76	1,542
75D	Personnel Records Specialists	34.07	65.93	41.86	58.14	989
76C	Equipment Records and Parts Specialist	94.17	5.83	58.34	41.66	2,403
76Y	Unit Supply Specialist	83.92	16.08	56.84	43.16	4,279
81E	Graphics Documentation Specialist	62.79	37.21	83.72	16.28	129
88H	Cargo Specialist	87.99	12.01	59.29	40.71	533
91A	Medical Specialist	83.41	16.59	72.96	27.04	1,790
94B	Food Service Specialist	81.04	18.96	46.79	53.21	3,787
95B	Military Police	86.91	13.09	92.11	7.89	2,369
Total (percent)		83.16	16.84	65.41	34.59	100.00
Total (N)		30,209	6,119	23,761	12,567	36,328

Table 2

Mean Prediction Errors (Fairness Measures) and Common Metric (CM) for PE Method

	Prediction Error Scores		t-test		CM, Percent Individuals Underpredicted	
	Female	Black	Female	Black	Female	Black
13M						
Mean	0.000	-0.250	0.000	-2.199*	00.0	72.7
SD	0.000	0.653				
N	0	33				
%	0.00	8.80				
16D						
Mean	0.125	-0.336	0.638	-3.137**	46.9	71.4
SD	1.105	0.749				
N	32	49				
%	11.47	17.56				
31K						
Mean	-0.016	-0.046	-0.260	-0.260	55.3	54.7
SD	0.888	0.902				
N	219	1083				
%	7.96	39.38				
31V						
Mean	0.320	-0.112	4.281**	-2.849*	34.6	58.0
SD	0.853	0.880				
N	130	505				
%	7.52	29.21				
33T						
Mean	-0.349	-0.665	-0.908	-0.000	33.3	100.0
SD	0.666	0.000				
N	3	1				
%	4.23	1.41				
45K						
Mean	-0.606	-0.015	-2.041*	-0.107	85.7	51.9
SD	0.786	1.026				
N	7	54				
%	2.13	16.46				
52D						
Mean	-0.153	-0.101	-2.038*	-2.884**	56.9	55.2
SD	0.786	0.796				
N	109	522				
%	4.55	21.80				

Notes

CM, Common Metric – percentage of minority individuals underpredicted

* Statistically significant at .05 level

** Statistically significant at .01 level

	Prediction Error Scores		t-test		CM, Percent Individuals Underpredicted	
	Female	Black	Female	Black	Female	Black
54B						
Mean	-0.189	0.065	-1.941	1.409	61.4	48.2
SD	0.889	0.794				
N	83	297				
%	7.70	27.55				
55B						
Mean	-0.176	-0.084	-2.477**	-1.603	64.6	56.5
SD	0.632	0.834				
N	79	255				
%	8.60	27.75				
63B						
Mean	0.230	-0.041	5.605**	-1.725*	40.9	53.9
SD	0.819	0.814				
N	399	1155				
%	8.99	26.02				
68J						
Mean	-0.230	-0.271	-1.220	-2.900*	58.3	75.0
SD	0.654	0.725				
N	12	60				
%	3.27	16.35				
71D						
Mean	-0.251	-0.058	-3.594**	-0.714	61.0	53.6
SD	0.917	0.846				
N	172	110				
%	31.27	20.00				
71L						
Mean	-0.153	0.023	-3.962**	0.485	61.5	54.1
SD	0.888	0.961				
N	527	407				
%	68.89	53.20				
71M						
Mean	-0.134	0.062	-1.759*	0.517	60.2	50.6
SD	0.861	1.072				
N	128	79				
%	33.95	20.95				
72E						
Mean	-0.131	-0.123	-1.805*	-2.201*	52.9	55.3
SD	0.845	0.910				
N	136	266				
%	21.32	41.69				

	Prediction Error Scores		t-test		CM, Percent Individuals Underpredicted	
	Female	Black	Female	Black	Female	Black
72G						
Mean	-0.120	0.014	-2.633**	0.249	58.5	51.0
SD	0.823	0.885				
N	325	261				
%	50.08	40.22				
73C						
Mean	-0.148	-0.024	-3.117**	-0.510	58.0	50.8
SD	0.887	0.898				
N	350	362				
%	43.80	45.31				
75B						
Mean	-0.184	-0.012	-4.599**	-0.364	62.5	52.4
SD	0.885	0.894				
N	491	721				
%	31.84	46.76				
75D						
Mean	-0.047	0.054	-1.277	1.311	53.5	50.4
SD	0.935	0.979				
N	652	575				
%	65.93	58.14				
76C						
Mean	-0.025	-0.072	-0.289	-2.833**	58.6	57.1
SD	1.021	0.803				
N	140	1001				
%	5.83	41.66				
76Y						
Mean	-0.266	-0.007	-8.312**	-0.301	65.8	55.1
SD	0.839	0.944				
N	688	1847				
%	16.08	43.16				
81E						
Mean	-0.217	0.150	-1.670	0.844	62.5	42.9
SD	0.899	0.815				
N	48	21				
%	37.21	16.28				
88H						
Mean	-0.163	-0.118	-1.506	-1.821	56.3	53.9
SD	0.864	0.953				
N	64	217				
%	12.01	40.71				
91A						
Mean	-0.230	0.036	-4.245**	0.852	61.3	49.0
SD	0.936	0.930				
N	297	484				
%	16.59	27.04				

	Prediction Error Scores		t-test		CM, Percent Individuals Underpredicted	
	Female	Black	Female	Black	Female	Black
94B						
Mean	-0.222	-0.040	-6.779**	-2.155*	65.0	55.3
SD	0.876	0.830				
N	718	2015				
%	18.96	53.21				
95B						
Mean	-0.179	0.099	-3.585**	1.388	62.6	47.1
SD	0.877	0.977				
N	310	187				
%	13.09	7.89				
Grand Arithmetic Mean	-0.128	-0.033				
Grand Absolute Mean	0.173	0.050				
Overpredictions	3	8				
Underpredictions	22	18				

The PE fairness score is computed for each individual and then the mean score is computed by minority subgroup within an MOS. Each MOS total sample has converted prediction scores with a mean equal to zero and a SD equal to one within the sample. Since the criterion (SQT) scores also were standardized within an MOS to have a mean of zero and a SD of one, this permits us to have equivalent scales for computing PE fairness measures. Thus, while the PE means in the total MOS sample are zero, the minority subgroups for each MOS may have negative or positive mean PEs.

It can be seen by the summary at the end of Table 2 that 22 of 25 MOS have mean PE fairness measures indicating underpredictions for females. Of these underpredictions, 16 mean PE fairness measures were found to be statistically significant for females at either the .01 or .05 level. Additionally, in 21 of 25 MOS, 50 percent or more of the female individuals in an MOS were underpredicted.

The six jobs with the highest levels of statistical significance for the PE fairness measure were all underpredicted in the female sub-sample. Females made up 17 percent to 68 percent of the total number employed in these jobs compared to an average of 10 percent female employment for all 25 jobs. It should be stressed that these six jobs with the largest negative statistical significance fall within job types considered traditionally female jobs in administrative and clerical areas, including MOS in Unit Supply, Food Service, Personnel Records, Legal, Medical and Personnel Administration. Only 3 of the 25 MOS had positive mean PE fairness measures, showing overprediction, in female sub-samples. The performance underpredictions reported here are not uncommon in the military, while underpredictions in the civilian literature are relatively uncommon.

Although the grand overall arithmetic mean PE for females is –0.128 and not of practical significance for determining minimum cutoff scores for selection, it is, however, important for classification or MOS job assignment when an optimization procedure is utilized. The underprediction of female performance, again, is of particular concern because it is most prevalent for traditional female jobs.

In Table 2, it can be seen that 18 of 26 MOS were underpredicted for blacks. Of the 18 MOS, 10 MOS show statistically significant t-test results. Also, in 23 of 25 MOS, 50 percent or more of black individuals were underpredicted in each MOS. However, unlike the findings for females, the fairness measures do not have a distinct pattern of job types.

Cleary Model for Fairness Measures

The Cleary fairness measure (CFM) is computed as the difference between two regression lines. For each MOS, the parameters (b and c) described in the Methods section are computed in the total MOS sample for one line and the parameters for the other line are computed in either the female or black sample of that MOS. The difference between the line with female or black subgroup parameters and the line with total group parameters produces a female or black CFM score for each individual.

From the results of Table 3 for the Cleary fairness measures, we find 22 of 25 MOS are underpredicted by female CFMs. These CFM results appear quite similar to those obtained for the PE models for female sub-samples (where a like number of MOS were underpredicted). Also in 21 out of 25 MOS, over 50 percent of the members of the female sub-sample were under predicted by the composite scores as measured by the CFMs.

From Table 4, we see underpredictions for blacks using the CFM in 12 of 25 MOS compared to underpredictions using the PE method in 18 of 25 MOS. These findings for blacks are not quite as similar to the PE results (as was true for females when the two methods are compared). However, the black CFM means computed in the sub-samples, in general, had high negative values and the PE based t-tests had statistically significant levels for the same administrative and clerical jobs as were found to be underpredicted when measured by PE fairness measures or CFMs. The pattern of fairness measure magnitudes for blacks, across different MOS, appears more diffuse – a pattern not as clearly discernible as fairness measures for females.

Table 3

Mean Fairness Measures and Common Metric (CM) for Females for Cleary Model

	Cleary Fairness Measure		CM, Percent
	Subgroup Mean	Total Sample Mean	Individuals Underpredicted
13M			
Mean	0.000	0.000	0.0
N	0		
%	0.00		
16D			
Mean	0.181	0.216	0.0
N	32		
%	11.47		
31K			
Mean	0.022	-0.001	18.7
N	219		
%	7.96		
31V			
Mean	0.380	0.376	0.0
N	130		
%	7.52		
33T			
Mean	-0.289	-0.349	100.0
N	3		
%	4.23		
45K			
Mean	-0.513	-0.426	100.0
N	7		
%	2.13		
52D			
Mean	-0.077	-0.057	81.7
N	109		
%	4.55		
54B			
Mean	-0.123	-0.196	100.0
N	83		
%	7.70		

Notes
CM, Common Metric – percentage of minority individuals underpredicted
Total Sample – Means computed on two different samples using same parameters

	Cleary Fairness Measure		CM, Percent Individuals Underpredicted
	Subgroup Mean	Total Sample Mean	
55B			
Mean	-0.124	-0.126	100.0
N	79		
%	8.60		
63B			
Mean	0.297	0.226	0.0
N	399		
%	8.99		
68J			
Mean	-0.200	-0.195	58.3
N	12		
%	3.27		
71D			
Mean	-0.195	-0.186	100.0
N	172		
%	31.27		
71L			
Mean	-0.140	-0.138	100.0
N	527		
%	68.89		
71M			
Mean	-0.108	-0.133	100.0
SD	0.861		
N	128		
%	33.95		
72E			
Mean	-0.101	-0.116	100.0
N	136		
%	21.32		
72G			
Mean	-0.120	-0.126	100.0
N	325		
%	50.08		
73C			
Mean	-0.144	-0.149	100.0
N	350		
%	43.80		
75B			
Mean	-0.151	-0.149	100.0
N	491		
%	31.84		

	Cleary Fairness Measure		CM, Percent Individuals Underpredicted
	Subgroup Mean	Total Sample Mean	
75D			
Mean	-0.024	-0.024	100.0
N	652		
%	65.93		
76C			
Mean	0.001	-0.028	44.3
N	140		
%	5.83		
76Y			
Mean	-0.249	-0.245	100.0
N	688		
%	16.08		
81E			
Mean	-0.180	-0.218	93.8
N	48		
%	37.21		
88H			
Mean	-0.122	-0.086	95.3
N	64		
%	12.01		
91A			
Mean	-0.198	-0.183	100.0
N	297		
%	16.59		
94B			
Mean	-0.182	-0.215	100.0
N	718		
%	18.96		
95B			
Mean	-0.112	-0.139	100.0
N	310		
%	13.09		
Grand Arithmetic Mean	-0.095		
Grand Absolute Mean	0.154		
Overpredictions	4		
Underpredictions	20		

Table 4

Mean Fairness Measures and Common Metric (CM) for Blacks for Cleary Model

	Cleary Fairness Measure		CM, Percent
		Total Sample	Individuals
	Subgroup Mean	Mean	Underpredicted
13M			
Mean	-0.211	-0.192	100.0
N	33		
%	8.80		
16D			
Mean	-0.263	-0.300	100.0
N	49		
%	17.56		
31K			
Mean	-0.013	-0.005	89.8
N	1083		
%	39.38		
31V			
Mean	-0.057	-0.097	94.3
N	505		
%	29.21		
33T			
Mean	0.000	0.000	100.0
N	1		
%	1.41		
45K			
Mean	0.059	0.030	0.0
N	54		
%	16.46		
52D			
Mean	-0.048	0.003	80.8
N	522		
%	21.80		
54B			
Mean	0.107	0.128	0.0
N	297		
%	27.55		

Notes

CM, Common Metric – percentage of minority individuals underpredicted

Total Sample – Means computed on two different samples using same parameters

	Cleary Fairness Measure		CM, Percent Individuals Underpredicted
	Subgroup Mean	Total Sample Mean	
55B			
Mean	-0.069	0.049	73.3
N	255		
%	27.75		
63B			
Mean	0.031	0.023	0.0
N	1155		
%	26.02		
68J			
Mean	-0.241	-0.181	100.0
N	60		
%	16.35		
71D			
Mean	0.041	0.112	42.7
N	110		
%	20.00		
71L			
Mean	0.044	0.047	0.0
N	407		
%	53.20		
71M			
Mean	0.101	0.092	0.0
SD	0.861		
N	79		
%	20.95		
72E			
Mean	-0.092	-0.089	100.0
N	266		
%	41.69		
72G			
Mean	0.023	0.024	0.0
N	261		
%	40.22		
73C			
Mean	-0.012	0.026	66.3
N	362		
%	45.31		
75B			
Mean	0.020	0.049	44.2
N	721		
%	46.76		

	Cleary Fairness Measure		CM, Percent
		Total Sample	Individuals
	Subgroup Mean	Mean	Underpredicted
75D			
Mean	0.080	0.106	5.22
N	575		
%	58.14		
76C			
Mean	-0.040	-0.055	98.6
N	1001		
%	41.66		
76Y			
Mean	0.011	0.014	.6
N	1847		
%	43.16		
81E			
Mean	0.227	-0.016	23.8
N	21		
%	16.28		
88H			
Mean	-0.093	-0.025	80.2
N	217		
%	40.71		
91A			
Mean	0.086	0.096	0.0
N	484		
%	27.04		
94B			
Mean	-0.008	0.007	72.85
N	2015		
%	53.21		
95B			
Mean	0.200	0.266	0.0
N	187		
%	7.89		
Grand Arithmetic Mean	0.004		
Grand Absolute Mean	0.039		
Overpredictions	13		
Underpredictions	12		

Rank-Order and Intercorrelations

Table 5A shows MOS rank-order results on fairness measures and CM indices for females for the PE and Cleary models. The four columns of rankings place a digit of one for the MOS with the highest negative value and a 25 or 26 for the MOS with the highest positive value.[7]

Table 5B shows the Pearson product moment intercorrelations for females among the rankings of the MOS in terms of the four variables. The correlation between the PE fairness measure and the CFM is .95, a very high correlation. Correlations among the other variables range between .54 and .60. The lowest correlation coefficient is .54 (between the two common metrics). This is an expected result partly attributable to the common metric itself – the percentage of individuals in each minority sub-sample. Many MOS were comprised of individuals with 100 percent being underpredicted.

Table 6A shows MOS rank-order results on fairness measures and CM indices for blacks for the PE and Cleary models. Table 6B shows the Pearson product moment intercorrelations among the variables for blacks. Again, the correlation of .90 between PE fairness and CFM is the highest correlation among the four variables. Correlations among the other variables range from .81 to .90.

Considering the intercorrelations among the two minority groups taken together, they can be characterized as moderately high to moderately. Most important is the very high correlation of .95 and .90 for females and blacks, respectively, between the PE and Cleary fairness measures. In this context, it should also be noted that the grand arithmetic means are -.128 and -.033 for females and blacks, respectively, for the PE model (see the end portion of Table 2) compared to -.095 and +.004 for the Cleary model (see the ends of Tables 3 and 4) – a mean difference of -.033 for females and -.029 for blacks. Again, these values indicate the closeness of the two fairness models despite differences in definition, scale and methods of computation.

[7] In the event of ties, the rule followed is to assign to all MOS that are tied the average of the ranks that they would have received had they not been tied (Adkins, 1965, p. 84).

Table 5A

Rank-Order of Two Variables each for PE and Cleary Models for Females

MOS	PE Fairness	PE CM	Cleary Fairness	Cleary's CM
16D	23	22	23	24
31K	22	19	22	22
31V	25	24	25	24
33T	2	25	2	8.5
45K	1	1	1	8.5
52D	14	17	19	19
54B	9	9	13	8.5
55B	12	4	12	8.5
63B	24	23	24	24
68J	6	15	4	20
71D	4	11	6	8.5
71L	15	8	11	8.5
71M	17	12	17	8.5
72E	18	21	18	8.5
72G	19	14	15	8.5
73C	16	16	10	8.5
75B	10	6	9	8.5
75D	20	20	20	8.5
76C	21	13	21	21
76Y	3	2	3	8.5
81E	8	7	8	18
88H	13	18	14	17
91A	5	10	5	8.5
94B	7	3	7	8.5
95B	11	5	16	8.5

Table 5B

Intercorrelations of Four Variables for Females

	PE Fairness	PE CM	Cleary Fairness	Cleary CM
PE Fairness	1.00	.60	.95	.55
PE CM	.60	1.00	.58	.54
Cleary Fairness	.95	.58	1.00	.59
Cleary CM	.55	.54	.59	1.00

Table 6A

Rank-Order of Two Variables each for PE and Cleary Models for Blacks

MOS	PE Fairness	PE CM	Cleary Fairness	Cleary's CM
13M	4	2	3	2
16D	2	3	1	2
31K	12	11	10	7
31V	7	4	7	6
33T	1	0	13	2
45K	16	17	20	21.5
52D	8	9	8	8
54B	24	23	24	21.5
55B	9	6	6	10
63B	13	14	17	21.5
68J	3	1	2	2
71D	11	15	18	14
71L	20	12	19	21.5
71M	23	20	23	21.5
72E	5	8	5	2
72G	19	18	16	21.5
73C	15	19	11	12
75B	17	16	15	13
75D	22	21	21	16
76C	10	5	9	5
76Y	18	10	14	17
81E	26	25	26	15
88H	6	13	4	9
91A	21	22	22	21.5
94B	14	7	12	11
95B	25	24	25	21.5

Table 6B

Intercorrelations of Four Variables for Blacks

	PE Fairness	PE CM	Cleary Fairness	Cleary CM
PE Fairness	1.00	.89	.90	.87
PE CM	.89	1.00	.84	.81
Cleary Fairness	.90	.84	1.00	.85
Cleary CM	.87	.81	.85	1.00

SUMMARY AND CONCLUSIONS

Summary

The major purpose of the present research was to compare the Prediction Error fairness measure (PE) and the Cleary fairness measure (CFM) across 26 MOS for female and black soldiers using (as prediction measures) the seven test LSE composites of the existing ASVAB and (as criterion measures) the SQT. Fairness is traditionally defined as the absence of underpredictions for the minority groups that are considered potentially susceptible to discrimination. The two models were compared for fairness in classification (i.e., over the full range of scores) and evaluated on the same robust Army database using a common metric. A double cross-validation design permitted unbiased estimates of prediction fairness for PE. It was concluded that for purposes of selection or setting standards the two models were roughly comparable, but the authors consider the PE model more precise and more objective because it takes into account individual scores rather than the difference between two regression lines.

The comparisons between the PE and CFM are summarized below:

There were a large number of individuals underpredicted in each MOS by the common metric. For the PE method, we found underpredictions of 50 percent or more for 21 of 25 MOS (females) and for 24 of 25 MOS (blacks). For the CFM method, we found underpredictions of 50 percent or more for 15 of 25 MOS (females) and for 13 of 25 MOS (blacks). The correlation between PE and CFM was .95 for females and .90 for blacks.

As noted earlier, Cleary evaluated black and white college student groups, using SAT scores (the predictors) and GPA (the criteria). Black students numbered 273 and white students over 2,000. The authors of the present study believe that it is fair to say that the minority students were very carefully selected on the basis of grades, motivation, and other relevant factors. In this academic context, Cleary found the black group regression lines higher than the white lines.

In contrast, the present study numbered over 12,000 blacks out of about 36,000 recruits. ASVAB weighted composites were used in 26 MOS as predictors and the SQTs as criteria. We compared the results from the PE and CFM methods on an Army sample. While the PE and CFM methods had different definitions and methods of computing fairness, we found roughly comparable results.

Conclusions

Based on the results and comparisons given above, the conclusions made are:

The large number of performance underpredictions for black and female soldiers found in the present study were relatively small in magnitude and are deemed of little practical importance in the selection process or for setting MOS cutoff scores, but they may have significant consequences for classification if and when classification is designed to optimize recruit assignment, and the underpredictions do have undesirable social implications.

With the removal of the "speeded tests" from the ASVAB – Numerical Operations (NO) and Coding Speed (CS) – in January 2002, the authors have found more MOS with underpredictions and more individual underpredictions in each MOS, as well as a significant loss of mean predicted performance in classification experiments (Zeidner, et al., 1998; 2004).

Future minorities may constitute as much as 50 percent of recruits in the Army. This is a much higher percentage than is expected in the other services. For this reason alone, the Army should assess the issue of fairness as far as practicable.

The authors argue that in both Cleary's fairness measure and t-test of statistical significance there is bias introduced in computing regression lines that are dependent on group membership and not directly based on test and criterion scores of individuals, regardless of gender or race. In the Cleary model, this leads to higher t-test and CFM results as reported earlier in the results section and in Appendix C. Bias consists of overlapping information in the subgroup means of the CFM variable. This overlap in turn inflates both the correlation between CFM and group membership and the size of the t-test. An additional souce of inflation of the t-test is due to the use of variables that are computed in the same sample as the regression line parameters (i.e., back sample inflation).

PE is also considered a better measure of fairness, holding the issue of the biasing effects aside, because measures are computed directly for individuals rather than in terms of the distance between regression lines. Mean prediction error differences are equal to the aggregation of differences between pairs of predicted criterion scores at the individual level, rather than in terms of under- or overprediction of biased regression lines.

RECOMMENDATIONS: TOWARD IMPROVED CLASSIFICATION

Consider that societal values dictate that predictor tests be fair to minorities while making personnel decisions.

Reintroduce Coding Speed (and possibly Numerical Operations) ASVAB subtests for the Army's use in classification (as the Navy is now doing for CS). This will immediately ameliorate the fairness issue for minorities (and significantly increase mean predicted performance (MPP) when classification is optimized).

REFERENCES

Adkins, D.C. (1965). *Statistics.* Columbus, OH: Merrill, Inc.

Cascio, W.F. (1991). *Applied Psychology in Personnel Psychology.* 4th ed. Englewood Cliffs, NJ: Prentice Hall.

Cascio, W.F., Outtz, J., Zedeck, S., & Goldstein, I.L. (1991). Statistical implications of six methods of test score use in personnel selection. *Human Performance, 4,*, 233-264.

Cleary, T.A. (1968). Test bias: Prediction of grades for Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115-124.

Gottfredson, L.S. (1998). Reconsidering fairness: A matter of sound and ethical priorities. *Journal of Vocational Behavior, 33*, 293-319.

Guion, R.M. (1991). Personnel assessment, selection, and placement. In M.D Dunnette and L.M Hough (Eds.) *Handbook of Industrial & Organizational Psychology.* Pp. 327-397. Palo Alto, CA: Consulting Psychologists Press, Inc.

Guion, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions.* Mahnay, NJ: Lawrence Erlbaum.

Jaeger, R.M. (Ed.) (1976). On bias in selection [special issue]. *Journal of Educational Measurement, 13,*3-99.

McLaughlin, D.H., Rossmeissl, P.E., Wise, L.L., Brandt, D.A., & Ming-mei Wang (October, 1984). *Validation of current and alternative Armed Services Vocational Battery (ASVAB) area composites.* ARI Technical Report 651. Alexandria, VA.

McNemar, Q. (1949). *Psychological Statistics*. New York: John Wiley & Sons.

Ostle,B. & Mensing, R.W., (1975).Statistics in Research. 3rd Ed. The Iowa State University Press/ Ames.

Scholarios, D.M., Johnson, C.D., & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 3,* 412-434.

Schmidt, F.L., & Hunter, J.E. (1974). Racial and ethnic bias in psychological tests: Divergent implications of two definitions of test bias. *American Psychologist, 29,* 1-8.

Thorndike, R.L. (1971). Concepts of cultural fairness. *Journal of Educational Measurement, 8, 63-70.*

Wise, L., Welsh, J., Grafton, F., Foley, P., Earles, J., Sawin, L., & Divgi, D. R. (1992). *Sensitivity and Fairness of the Armed Services Vocational Aptitude Battery (ASVAB) Technical Composites* (DMDC Technical Report 92-002). Monterey, CA: Defense Manpower Data Center.

Zeidner, J., Johnson, C.D., Vladimirsky, Y., & Weldon, S. (September 1998). *Fairness of Proposed New ASVAB Test Composites for Restructured Job Families.* Submitted to ARI for publication. Alexandria, VA.

Zeidner, J., Johnson, C.D., Vladimirsky, Y., & Weldon, S. (August 2000). *Specifications for an Operational Two-Tiered Classification System for the Army, Volume 1: Report.* ARI Technical Report 1108. Alexandria, VA.

Zeidner, J., Johnson, C.D., Vladimirsky, Y., & Weldon, S. (April 2004). *Fairness of New Army ASVAB Test Composites for MOS and Job Families.* ARI Study Note 2004-04. Alexandria, VA.

Zeidner J., Scholarios, D., & Johnson, C.D. (2003). Evaluating job knowledge criterion components for use in classification research. *Military Research, 15*, 97-116.

APPENDIX A

COMPUTING STANDARDIZED, MOS SPECIFIC, COMPOSITE SCORES

Part One

Discussion of Basic Concepts

Summary

The process of obtaining composites consisting of best weighted ASVAB tests converted to standardized composite scores for the youth population (YP) begins with multiplication of operational test scores (1 by 7 row vector) and raw score regression weights (7 by 1 column vector), and adding a constant for each individual. The second step is to convert these preliminary composite scores to have a mean of 100 and a standard deviation of 20 in the YP. After this conversion to the Army standard score scale for the youth population (YP), these composite scores are expressed in the same scale as both the minimum cut scores that are provided for each MOS, and the nine best weighted "interim" composites that correspond to the nine operational job families. Either of these sets of composites, the nine or the 150 are second tier selection variables, as described in the next section. Either would be appropriate for use as the predictor variables in the algorithms for computing the Cleary fairness measures that are compared to the PE fairness measure in this study. Also, either of these sets of composites is also appropriate for computing the Thorndike fairness measures that are described but not included in the analyses of this study.

The predictor variable used in the algorithm for computing the PE fairness measure is obtained by similarly computing best weights to be applied to test scores in each MOS level job family, but correcting the components used to compute the regression weight parameters to the Army input population (AIP) instead of to the YP. The resulting PPs are then converted to statistical standard scores based on the means and standard deviations of each of the 150 MOS level job families.

First Tier Composites

The approximately 250 Army MOS have been placed in 150 first tier job families, most of which consist of single MOS families, but a few are clusters of low density MOS.

The first tier composites proposed for use as black box classification variables in a two tiered selection and classification system rely on the use of separate PPs for each MOS level job family. These PPs use separate weight vectors and regression constant for each of these 150 job families. The 7 best weights and the regression constant for each first tier composite are computed using validity coefficients computed against 150 MOS criterion variables (i.e., SQTs) corrected for attenuation and then for restriction in range to the Army input population (AIP). The variances of both tests and MOS SQT are also corrected to the AIP. These first tier composites have statistical standard scores standard deviations equal to the validity coefficients, corrected to the AIP, of each PP against the SQT for that MOS. This set of first tier composites has a potential, when used as the classification variables in an optimal assignment system, of greatly increasing the predicted performance of new soldiers. These first tier composites are converted to statistical standard scores, means equal to zero and standard deviations equal to 1.0, for use in computing PE fairness measures. This conversion removes the hierarchical classification effects property that are possessed by the first tier composites, reducing their classification efficiency by approximately ten percent.

This reduction in classification efficiency was necessary to permit both the predictor and criterion variables to be expressed as statistical standard scores.

Second Tier Composites

The 250 Army MOS are currently clustered for operational use into nine (9) operational job families, and separate best weighted composites (PPs) computed for each of these job families. These nine families and the second tier composites were not used in the present study, thus the following description is for background purposes only. Best weighted second tier composites are being used as replacements for the integer weighted aptitude area (AA) composites. These composites use optimal weights obtained through a process that starts with obtaining validity coefficients computed against the same 150 MOS criterion variables (i.e., SQTs) corrected for attenuation and then for restriction in range to the youth population (YP), and on variances also corrected to the YP. These 150 validity coefficients are then aggregated into validity coefficients representing the 9 official Army job families. The best weights are applied to the 7 tests and the resulting PP scores are then converted into Army standard scores in the YP.

The same procedure, involving u and k values applied to test scores as is used for computing the set of 9 second tier composites, is used in this study to compute a set of 150 composites corresponding to each of the job families for which the data is adequate to compute regression weights. It should be noted that these sets of 9 operational and 150 composites used to compute PE fairness measures, and the official minimum cut scores for each MOS are in the second tier scale. All three measures have been converted to the Army standard score scale in the YP.

Regression Weights For Composites

Regression weights applied to tests form PP scores that maximize the prediction of the SQTs utilized as the criterion variables in this study. SQT scores for 150 job families (mostly single MOS) were converted to statistical standard scores (SSSs) within the MOS samples. The above PP scores for each individual were similarly converted to SSSs and product moment correlation coefficients computed between these PP scores and the SQT SSSs separately in both halves of each MOS sample (in conformance with a double cross validation design).

The inter-correlation coefficients among the 7 ASVAB tests and both the criterion and predictor variances were also computed in job family half samples. These first tier job families at the MOS level were used in this study.

The test validity coefficients computed in the MOS level samples were corrected for unreliability of the criterion and for restriction in range effects to either the YP or AIP. Thus, both the variances of the predictor and criterion variables were corrected to either the YP or AIP and regression weights computed. The restriction in range correction procedures are thoroughly described in Appendix B.

Conversion of Best Weighted Composites to Standard Score Scale

Tier two composites, the only composites used in depicting the Cleary measures in this study, are PP variables corrected for restriction in range to the YP and then converted to have a mean of 100 and a standard deviation of 20 in the YP. This same composite is used in describing the Thorndike method in the methods section. The formula required to make this conversion of a preliminary "best weighted" composite is provided in the following section. Tier one composites, as described in previous studies, are PP variables corrected for restriction in range to the AIP and then converted to have means of zero and standard deviations equal to the validity coefficients against the MOS SQT. The predictor variables used to compute PEs are obtained by converting the tier one composite scores to statistical standard scores in each MOS.

Part Two

Notation, Formulae, and Use of "u and k" parameters with Operational Test Scores: Application of Basic Concepts

Introduction

The discussion of notation must precede a discussion of formulae, followed by a discussion of algorithms, one of which utilizes the u and k parameters that are applied to operational test scores to obtain best weighted test composites used in this study. The presentation of notation will begin with the definition of subscripts that identify the degree or lack of restriction in range correction, the population to which conversion efforts lead, the MOS to which best weights apply, and indicate whether individuals or groups provided score vectors. These subscripts are applied to matrices, vectors or scalars.

Notation

The subscript categories: g, h, i, j, or the numbers identifying the sub-category within the category, may be attached to the matrix, vector, etc. in alphabetical order. The meaning of subscript categories and the numbers identifying the sub-categories are listed below:

a. The sub-categories of g identify the level of restriction in range correction pertaining to the symbol to which the subscripts are attached; g = 0 indicates no such correction is utilized; g = 1 indicates that corrections for restriction in range have been made to the Army input population; and g = 2 indicates that these corrections have been made to the youth population. The computation of a value in a sample drawn from the indicated population can be substituted for a correction to that population.

b. The subcategories of h identify the population to which a conversion in scale has been made: h = 0 indicates that no conversion in scale is made; h = 1 indicates the conversion is to the first tier scale (M = 0, SD = validity coefficient for predictor); h = 2 indicates the conversion is to the second tier scale (for composites, M = 100, SD = 20).

c. The presence of the subscript 'i' indicates that there is a separate value for each individual and "j" indicates that there is a separate value for each MOS (i.e., job family). The i and/or j subscript will be omitted when it should be clear to the reader that computations are accomplished separately for each individual and/or MOS. Only scalar values or 1 by 7 vectors have "i" or "j" subscripts, that is, no matrices have "i" or "j" subscripts.

The above subscripts are attached to scalar variables, vectors, or matrices. The relevant subscripts are shown in the following definitions:

T_i = 1 by 7 vector of ASVAB test scores for each individual;

Y_i = 1 by 7 vector of SQT criterion scores for each individual;

W_{ghj} = 7 by 1 vector of regression weights used in the prediction of SQT scores;

z_{ghj} = a scalar number representing the individual LSE scores of a specified SQT variable;

R_{ghj} = a 7 by 7 matrix whose elements are correlation coefficients among the ASVAB tests corrected to, and/or obtained in the indicated population;

V_{gj} = a 1 by 7 vector whose elements are validity coefficients between the predictors and the SQT criterion for the jth MOS;

M_{tgh} = a 1 by 7 vector whose elements are ASVAB test means;

m_{zgj} = a scalar number representing the mean of a predictor (a composite of "best" weighted ASVAB tests) for the jth MOS;

m_{ygj} = a scalar number representing the mean of SQT scores for the jth MOS;

S_{tg} = a 7 by 7 diagonal matrix whose diagonal elements are standard deviations of tests; these elements may be defined for a specified population, or corrected for restriction in range to that population;

s_{zgh} = a scalar number representing the standard deviations of the composite of best weighted ASVAB test scores as defined in the population to which the scale conversion process is targeting;

t_i = operational ASVAB test score of an individual;

z_{gh} = the least square estimate (LSE) of the criterion (Y);

u_{gh} = multipliers for x gh in the process of converting x_{gh} to z_{gh};

k_{gh} = regression constant used in the process of converting x_{gh} to z_{gh}.

Basic Formulae

(1a.)$W_{ghj} = (R_{ghj})^{-1} (V_{ghj})'$; regression weights for application to predictor scores in statistical standard score format; a 7 by one vector of weights.

(1b.) B = a 1 by 7 vector for which each element is equal to the ratio of the SD of the criterion scores divided by a SD of an ASVAB test (each element is equal to (s_{yg} / s_{xg} multiplied by the corresponding standard score regression weight. These elements are the raw score regression weights. Note that each criterion SD is equal to 1.0 within a MOS for which the criterion scores have been converted to statistical standard scores (SSSs).

(2.) $z_{ghj} = T_i\, B_{ghj} + (m_{ygj} - M_{tghj}\, B_{ghj})$; predicted performance scores that can also be described as LSEs of the criterion.

(3.) $q_{ghj} = 2 / \{ (W_{ghj})'\, R_{gh}\, W_{ghj})^{1/2}\}$

A 1 by 7 vector of q weights is designated as Q.

(4.) $x_{ghj} = Q_{ghj} \ z_{ghj}$; composite scores used to compute both Cleary and Thorndike fairness measures, when both g and h equal 2. These scores (g and h = 2) are in the same scale as the minimum cut scores determined by an Army panel.

(4.a) $u_{ghj} = \ \{Q_{ghj} \ W_{ghj} (s_{yghj} / s_{tghj})\}$, u weight for the jth operational test score. A 7 by 1 vector of u weights is designated as U_j .

(4b.) $k_{ghj} = m_{yghj} - M_{tghj} \ U_j + 100$

(4c) $x_{ghj} = (T_i \ U_j) \ + k_{ghj}$

Discussion of Key Formula

The PP scores as defined in formula (2), when g = 1 and h = 0 are the first tier composites proposed for use in determining first tier minimum cut scores and in effecting optimal assignments. The PP scores that are converted to statistical standard scores for use in computing the PE fairness scores are also defined in formula (2) with g = 2 and h equal to zero. A composite in the first tier scale has a mean of zero and a standard deviation within each MOS equal to the composite's validity coefficient.

The composite scores used in computing CFM and in describing the Thorndike fairness method are as defined in formula (4c), when both g and h equal 2. These latter composite scores are converted to the second tier scale which makes them comparable, with respect to scale, to the operational test scores and to the minimum cut scores provided by the Army panel. Composite scores in this second tier scale have a mean of 100 and a standard deviation of 20 in the youth population.

The u values applied to test scores are obviously more than regression weights since they also convert the PP scores to Army standard scores in the YP. Since the operational tests have a mean of 50 and a standard deviation of 10 in the YP, and have moderately high inter-correlation coefficients, the conversion of a weighted test composite to the YP with a mean of 100 and a SD of 20 requires the multiplication of a weighted test sum by Q.

APPENDIX B

IMPACT OF RESTRICTION IN RANGE ON THE ESTIMATION OF AA COMPOSITES

Introduction

This appendix will focus on how to obtain the AA composite regression weights (referred to as "u and k" values) for operational use in the applicant Youth Population.[8] The validity coefficients we wish to maximize in the Youth Population actually exist only in doubly restricted MOS samples containing the Skill Qualifications Test (SQT) criterion in the 1987 - 1989 research data set. Appropriate corrections have to be made to these restricted validity coefficients to obtain unrestricted validity coefficients that, if subjected to restriction in range effects, would equal what was obtained in the MOS samples. We also have to estimate what the criterion standard deviation (SD) would have to be in the unrestricted population to yield the criterion SDs observed in the MOS samples.[9]

The Army operational process involves an applicant Youth Population from which self-selection first occurs, and then the Recruiting Command selects some and rejects others using tests, medical examinations, security investigations etc. This results in an Army Input Population from which classification and assignment procedures and further self selection create the 150 MOS samples, each with its separate SQT criterion measure. Thus there is a selection stage and a classification and assignment stage, with a restriction in range effect on both test scores and hypothetical criterion scores occurring at both stages. If we confined selection effects to the impact of the AFQT screen, the two kinds of effects would have to be corrected in a sequential manner. However, since we are not restricting ourselves to such a limited selection

[8] This appendix has been prepared by Cecil Johnson, consulting research psychologist.

effect, and are instead considering all effects on the subtest co-variances at each restriction stage, we can correct validity coefficients and criterion SDs directly to the Youth Population.

Our correction process for restriction in range involves contrasting, separately for each MOS, the within-MOS subtest variance/co-variances against the Youth Population operational test variance/co-variances. The differences in the variance/co-variances across the unrestricted and the restricted samples for variables specified as explicitly selected variables are the measures of the magnitude of the restriction effect. For our purposes we use all ASVAB subtests as the explicitly restricted variables and we designate the criterion variables as the implicitly restricted variables that are restricted to the extent that they are predicted by the explicitly restricted variables.

Using this concept we can calculate the effect selection has on subtest scores and can then calculate the further effect classification and assignment has on test scores in the Army Input Population – to arrive at the doubly restricted subtest scores in the MOS samples. Considering the correlation of the subtest scores with the criterion scores and the amount of restriction occurring at each stage, we can determine the restriction effect on the hypothetical criterion scores and then provide a correction extending from the MOS criterion scores to the less restricted populations where the criterion scores exist only as a function of the subtest scores (i.e., as predicted criterion scores).

Approach

There is more than one algebraically equivalent way of providing operational u and k values when criterion scores are only available on the doubly restricted MOS samples. We will

[9] It should be noted that whenever validity coefficients are mentioned, we are assuming that these coefficients have been corrected for attenuation with respect to criterion unreliability. Even if we should refer to an uncorrected validity coefficient (for restriction in range), this "uncorrected" coefficient has been corrected for attenuation.

use an approach that utilizes the equality of G-weights computed in the restricted and the unrestricted population (using Gulliksen's formulation as described below). The G-weights computed in the restricted population samples will be used as a substitute for the unobtainable G-weights in the unrestricted population in Gulliksen's formula for computing the criterion variance in the unrestricted population.

1. Consider the matrix of G-weights, G, in each MOS sample. Our use for G is as an entry value in Gulliksen's formula (see below). The corrected validity coefficients, obtained with the use of the formula at either or both the Army Input Population and Youth Population points, were then employed in computing Beta weights in the Youth Population. Note that this correction must be made from each MOS sample to the Youth population to produce validity coefficients corrected for restriction in range. These corrected MOS validity coefficients are then aggregated into a corrected validity for each specified family, using acquisition values to weight the MOS validity coefficients corrected to the Youth Population.

2. Visualize a composite computed for an individual by summing the product of each subtest standard score and B. The best weighted composite XB will have a SD equal to the validity of predicted performance (PP) in the Youth Population if the elements of the V matrix used in computing B are validity coefficients corrected for restriction in range to represent the Youth Population, and the R matrix consists of the inter-correlation coefficients among subtests as expected in the Youth Population. The criterion variables, predicted as least square estimates (LSEs) by the PP composites, have a SD equal to 1.0 in the restricted MOS samples, while the hypothetical unrestricted criterion variables would have larger SDs in the less restricted populations. Compute the Youth Population beta weights as follows:

   $$B = R^{-1} V^{T},$$

   where R is the Youth Population matrix of subtest inter-correlation coefficients and V is the matrix of validity coefficients corrected to the Youth Population. Looking at the formula in more detail,

   $$R = S_x C_{xx} S_x, \text{ and } V^{T} = S_x C_{xc} S_{c,}$$

   where C represents criterion / subtest variance and co-variances found in Gulliksen's formulae, and S represents a diagonal matrix where each diagonal element is equal to a reciprocal of a SD.

3. Compute b-weights by converting the Beta weights computed in step 2. The b-weights that are appropriate to apply to operational test scores to obtain a least squares estimate (LSE) of the criterion can be defined in terms of the Beta weights, the SDs of the subtests, and the SDs of the criterion scores. These b-weights applied to the operational test scores would provide a composite that, if the appropriate regression constant were subtracted, would have a mean of 50 and a SD less than 10 (because of the effects of the positive inter-correlation

coefficients among the subtests). The b-weights are computed, ignoring the regression constants, as follows:

$$\text{b-weight} = \text{B-weight} * (SD)_c \; / \; (SD)_t \,,$$

where t represents a subtest, $SD_t = 10$, and c represents the criterion variable.

4. The composite computed in step 3 will have a SD less than 10. We wish to convert this composite to have a SD of 20. To do this we will multiply each b-weight by a composite multiplier (CM) that will convert the composite to have a SD of 20 without affecting the composite mean. CM can be computed as follows.

$$CM = 20 \, / \, (10 * (\underline{b}R\underline{b}^T)^{1/2}),$$

where $\underline{b}$ is a vector of b-weights and R is the Youth Population matrix of subtest inter-correlation coefficients.

5. We can now compute the u and k values for each composite:

$$u_j \;=\; CM * \text{b-weight of the j-th subtest}$$
$$k = \; 100 - \; \sum u_j * 50$$

Key Formulae From Gulliksen

The algorithms we use to correct for restriction in range due to "selection" effects are developed and described by Gulliksen (1950)[10]. His development is based on a model that visualizes the presence of both explicit and implicit selection processes in the unrestricted population, and the presence of both explicitly and implicitly selected variables in the restricted population. Thus, both explicit and implicit variables are present in both the unrestricted and restricted populations. The author shows, in the context of this model, relationships among the restricted and unrestricted variances/co-variances without relaxing flexibility as to which population contains the unknowns that cannot be directly computed but can be determined on the basis of the relationships defined in his model.

The Gulliksen formulae for correcting variances and/or co-variances for restriction in range effects are based on Lawley's (1943) assumptions that include the following: (1) that the

[10] See H. Gulliksen, *Theory of Mental Tests*. New York: John Wiley & Sons, 1950.

regression of the implicitly restricted variables on the explicitly restricted predictors is linear; (2) that the co-variance of the restricted variables exhibit homoscedasticity; and (3) that the G-weights for application to the population variance-covariance matrix of operational test scores (explicitly restricted variables, e.g., sub-tests) are invariant to the effects of restriction (as defined). Thus it is assumed that

$$G = (C_{xx})^{-1} \; (C_{xc})^{T}$$

can be computed in a restricted population sample and substituted in formulae for use in the unrestricted population where a G-weight is to be entered. Gulliksen's formula 42, used to compute criterion variance in the Youth Population, requires such an entry. This criterion variance is essential for converting Beta-weights into b-weights and obviously cannot be directly computed in the Youth Population.

As previously stated, our objective is to have an algorithm replete with valid formulae that will convert operational test scores into LSEs of the criterion (i.e. PP composites) in a scale appropriate for use in the indicated population.

Application of Formulae 37 and 42

Applying combined formulae 37 and 42 to one criterion variable at a time, and making small changes in Gulliksen's notation, we can compute the squared SD of each Youth Population criterion variable associated with each job family. This result can be described as the Youth Population criterion variance, or YPCV:

$$\text{YPCV} = 1.0 + \underline{C}_{xc} \, (C_{xx})^{-1} \, (\, (\, {}^{*}C_{xx}) \, (C_{xx})^{-1} - I \,)(\, \underline{C}_{xc})^{T},$$

where ($\underline{C}_{xc}$) is a 9 by 1 vector of co-variances between the criterion variable and each of the 9 tests, C_{xx} is a 9 by 9 matrix of co-variances among 9 tests using the operational test scores, and vectors are denoted by underlining. Note that the asterisk matrix, e.g. *C, indicates computation in the unrestricted (i.e. Youth Population) sample.[11]

The R matrix has the following relationship with the C_{xx} matrix:

$$R = S_x C_{xx} S_x,$$

where S is a diagonal matrix for which the diagonal elements are equal to the reciprocals of the SDs of either the subtests or the criterion variable in either the MOS sample or the Youth Population, as indicated.

The $*C_{xc}{}^T$ matrix is derived from the Gulliksen formula as:

$$(*C_{xc})^T = (*C_{xx})\ (G) = (*C_{xx})\ (C_{xx})^{-1}(C_{xc})^T \ .$$

Note that one column of $*C_{xc}{}^T$ is ($\underline{C}_{xc}$)^T , a vector used in the computation of YPCV. The validity matrix ($*V^T$) required to compute Beta weights in the Youth Population has the following relationship with the $*\underline{C}_{xc}{}^T$ vector:

$$\text{one column of } *V^T \text{ is } (*S_x)\,(*\underline{C}_{xc})^T\ (*\underline{S}_c)\ ,$$

and note that $*\underline{S}_c$ is a scalar.

Positively Weighted Composites for the Visible Tier

This section extends the initially professed objectives of this appendix beyond restriction in range corrections and the conversion of Betas to u and k values. We will now discuss the

[11] Note that YPCV can also be written as follows:

$$YPCV = 1.0 + (W^T)(*C_{xx}\,W - (\underline{C}_{xc})^T\ ),$$

where $W = (C_{xx})^{-1}\ (\underline{C}_{xc})^T$, a 9 by 1 vector of regression weights for a specified job family. W will also be recognized as one column of the G matrix.

methodology for selecting the "best" positively weighted composites where best is defined in terms of maximizing the multiple correlation coefficient of a set of tests with the criterion.

The surest way to find this best positively weighted composite from a set of n tests is to compute the Betas and validity coefficients for every possible combination of n tests, then successive levels: for n-1 tests, then n-2 tests, …to 2 tests --- rejecting any combination of tests that has one or more negative weights. There is no need to actually consider all of these combinations since there comes a point in this process where all multiple correlation coefficients (Rs) for succeeding levels are lower than the highest R in a prior level.

The multiple-correlation coefficient, R, corresponding to each set of Betas is computed for each combination whether or not all of the weights are positive. Clearly, if the R for each combination of m-1 tests, negative weights permitted, was less than the highest R for m positively weighted subtests computed from the combinations considered at the prior level, the stopping point has been reached. After the stopping criterion has been reached, the set of subtests with all positively weighted coefficients that provides the maximum R is selected as the very best set and these weights become the B-weights for the associated subtests. All other tests are given a weight of zero in the composite associated with the specified job family.

APPENDIX C

FAIRNESS MEASURES AS PREDICTORS OF GROUP MEMBERSHIP

Introduction

This appendix explores the properties of the CFM in a special context. We first consider a t-test that could be used to assess the statistical significance of the difference obtained by subtracting the CFM mean computed in the total MOS sample , used by us as a surrogate for the population corresponding to the MOS sample. Our concern is with the extent the mean black (or female) CFM in the minority sample is inflated in the minority sub-sample (a back sample regarding those parameters computed in the minority sample)—and inflated in the opposite direction in the total sample from using "back sample" parameters computed in the total MOS sample. These two separate biasing effects are aggregated in computing the differences between the sub-sample and total MOS sample scores that constitute the t-test formula of the t-test described above.

We will also compare the t-test formula to a bi-serial correlation formula where the continuous variable is CFM and the dichotomous variable represents membership in one of two alternative groups (e.g., black vs. white, female vs. male). It will be shown that the correlation of the CFM scores with group membership is inflated as compared to the correlation of PE scores, or other scores computed without knowledge or consideration of an individual's membership in any group. We then show the effect a change in the magnitude of the correlation between the fairness measure and group membership has on the magnitude of the t-statistic. In conclusion, we discuss why the CFM is certain to have a considerably higher correlation with group membership, as compared to when PE is used to compute this bi-serial correlation , in addition to the inflation due to back validity effects.

Formulae

The following formulas apply to black, white and total groups , but the same concepts behind these formulae can be applied to female, male and total groups. All formulae under sets 1 through 5 apply to either PE or Cleary fairness measures.

1. Numerator of biserial r = $(M_b - M_w) P_b (1 - P_b)$; P_b is the percentage of total MOS sample that is black., M_b is mean of fairness measures in black sub-sample, and M_w is the mean of fairness in the white sub-sample; $P_w = (1 - P_b)$.
Numerator of $t = (M_b - M_t) = (M_b - P_b M_b - P_w M_w)$;
where $P_w = (1 - P_b)$and M_t is the mean of fairness measure sample in the total MOS sample.
Thus, $M_t = P_b M_b + (1 - P_b) M_w$,
and numerator of $t = (M_b - M_w) (1 - P_b) = (M_b - M_t)$

2. Biserial r $= \{(M_b - M_w) (1 - P_b) P_b\} / \{(SD)_t (z) \}$;

Where $(SD)_t$ is the standard deviation of the fairness measure in the total MOS sample, and z is the ordinate on the normal curve at the point where the tail has an area equal to the smaller of P_b or $(1- P_b)$.

3. $t = [(1-P_b) (M_b - M_w) / (SD)_b] (N)^{1/2}$; where (SD) b is the standard deviation of a fairness measure in the black sample.

4. Expressing "t" by a function which includes "r " as a multiplier:

$t= r [\{ SD)_t / (SD)_b \} \{ z / (P_b)\}\{(N)^{1/2}\}]$

5. Expressing "r" by a function which includes "t" as a multiplier:
$r = t [\{ (SD)_b / (SD)_t \} \{ P_b / z (N)^{1/2} \}$

Further Discussion and Conclusions

A fairness measure can have a positive or negative correlation with membership in a minority group. A positive correlation may reflect characteristics often present in individuals in minority groups, such as poor test taking ability as compared with his/her on-the-job capability. In another type of MOS for which written tests are relatively low predictors of job performance the test taking ability may exceed performance capability providing a negative correlation (e.g.

for females in MOS requiring skills seldom taught to females). Both t-tests and bi-serial correlation coefficients would be negative in this situation.

The above characteristics, without further contamination, would not constitute a dependence between the fairness measure and the dichotomous variable described above. However, if the continuous variable and the dichotomous variable have a built in dependence, the t tables based on Students distribution would not be appropriate for use with the obtained values of t. We found such a dependence to be present for t-tests computed using the CFM ,and although these values were computed they are relegated to an appendix (Appendix D) and are not discussed in the main body of the report.

APPENDIX D

T-TESTS FOR CLEARY MODEL FOR FEMALES AND BLACKS

	t-test	
MOS	Female	Black
13M	0.000	-3.373**
16D	2.716**	8.128**
31K	13.583**	-31.150**
31V	14.511**	22.575**
33T	0.494	0.000
45K	-1.221	8.789**
52D	-2.208*	-20.182**
54B	8.192	-16.198**
55B	0.573	-13.901**
63B	26.246**	48.097**
68J	-0.051	-7.024**
71D	-5.182**	-6.238**
71L	-3.003**	-7.184**
71M	5.305**	8.347**
72E	5.730**	-12.326**
72G	3.052**	-6.641**
73C	2.168*	-9.776**
75B	-7.142**	-14.081**
75D	6.231**	-8.575**
76C	5.501**	19.293**
76Y	-6.573**	-28.960**
81E	1.887*	4.269**
88H	-4.358**	-10.100**
91A	-7.855**	-27.219**
94B	16.525**	-36.049**
95B	12.590**	-17.605**

Notes
** Statistically significant at .01 level
* Statistically significant at .05 level

Appendix E

The Effect on the Cleary Fairness Measure of the Proportion of Minorities in the Total Sample

Introduction

We will establish in this appendix, under the condition that the population means of the predictor and criterion variables remain constant, that the expected magnitude of the mean Cleary Fairness Measure (CFM) for a minority group decreases (i.e., overprediction decreases) as the ratio of the minority group size to the total group size increases. Thus, an increase in the proportion of minorities in the total group increases the expectation of underprediction in the minority group when the CFM is used to measure fairness. For example, this makes the finding of underprediction in the Army more likely than in the Air Force. However, when the mean predictor score in the minority group exceeds the mean predictor score in the total group, underprediction in the minority group requires that the mean criterion score in the minority group be negative (i.e., less than the mean criterion score in the total group, since the latter is set at zero in this study).

Notation

The means of either the predictor variable (x) or the criterion variable (y) will be expressed in bold caps, as: $\mathbf{X}_g$ for the predictor mean in a minority group; $\mathbf{X}_t$ for the predictor mean in the total MOS group, $\mathbf{Y}_g$ for the criterion mean in the minority group, and $\mathbf{Y}_t$ for the criterion mean in the total MOS group.

The Cleary Fairness Measure can be defined in terms of the difference between two regression equations:

$$\mathrm{CFM} = [b_g x + C_g] - [b_t x + C_t].$$

If the mean CFM for the minority group is denoted as $\mathbf{(CFM)}_g$, the CFM as measured in the minority group sample can be written in terms of predictor and criterion means as follows:

$$\mathbf{(CFM)}_g = [b_g \mathbf{X}_g + (\mathbf{Y}_g - b_g \mathbf{X}_g)] - [b_t \mathbf{X}_g + (\mathbf{Y}_t - b_t \mathbf{X}_t)] .$$

Since the mean of the first bracketed term is equal to the criterion mean in the minority group (i.e., $\mathbf{Y}_g$), the mean (CMF)g can be expressed as:

$$\mathbf{(CFM)}_g = \mathbf{Y}_g - [b_t \mathbf{X}_g + (\mathbf{Y}_t - b_t \mathbf{X}_t)].$$

Algebraic Simplification and Interpretation

Noting that: (1) $\mathbf{Y}_t$ is zero by definition in this study, since the criterion scores are all converted to statistical standard scores within each total MOS sample, and (2)$\mathbf{Y}_g$ is the mean of $b_g\,x + C_g$, a simplification process yields a formula for $\mathbf{CFM)}_g$ written as follows:

$$(\mathbf{CFM})_g = \mathbf{Y}_g - b_t\,(\mathbf{X}_g - \mathbf{X}_t)\,.$$

With $\mathbf{Y}_g$ and b_t specified to be invariant, along with the mean of x in the minority and non-minority sub-groups of a specified MOS population, we now consider the effect of varying the proportion of the total sample drawn from the minority population. Under these conditions the difference between mean of x found in the two sub-samples, minority and total, will be maximized as the minority sample size approaches zero and will be minimized as the minority sample approaches becoming 100 percent of the total sample. The smaller this difference, the larger CFM becomes with $(\mathbf{CFM})_g = \mathbf{Y}_g$ when $\mathbf{X}_g = \mathbf{X}_{t.}$.

Conclusions

It is potentially misleading to compare CFM values across studies where the proportions of minority individuals for a specified job widely differ. This would be particularly striking when comparing fairness results across comparable technical jobs in the Army and Air Force, because of the relatively low percentage of minorities in the Air Force.